EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Institute for the Protection and Security of the Citizen

# Sensitivity analysis: how to detect important factors in large models

**Francesca Campolongo and Jessica Cariboni**

**Institute for the Protection and Security of the Citizen**

**2007**

**EUR 22798 EN**

European Commission
Directorate-General Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
E-mail: Jessica.cariboni@jrc.it

http://finecon.jrc.it
http://www.jrc.cec.eu.int

Legal Notice
Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

ISSN 1018-5593
EUR22798 EN
Luxembourg: Office for Official Publications of the European Communities

© European Communities, 2007

Reproduction is authorised provided the source is acknowledged

Printed in Italy

# Sensitivity analysis: how to detect important factors in large models

*Francesca Campolongo and Jessica Cariboni*

European Commission, Joint Research Centre, Ispra (VA), Italy

April 2007

**Abstract**

In this work we propose a sensitivity analysis method which is effective in identifying the few important factors in a model that contains many factors, with a relatively small number of model evaluations.

The method is convenient when the number of factors is so large, and/or the model execution time is such, to make the computational cost of more sophisticated techniques excessive.

The method is conceptually simple and can be though as an expansion of a derivative-based approach, although it overcomes the limitation of a local derivative-based approach as it attempts to explore the whole input space.

## Table of contents

## Introduction

In this work we describe a sensitivity analysis method which is effective in identifying the few important factors in a model that contains many factors, with a relatively small number of sample points properly distributed.

The method is conceptually simple and easy to implement. Sticking to the concept of local variation around a base point, this method makes an effort to overcome the limits of the derivative-based approach by introducing wider ranges of variations for the inputs and averaging a number of local measures so as to lose the dependence on a single sample point.

The use of this method is ideal when the number of input factors is too large to allow the application of the computationally expensive variance-based techniques, but at the same time not so large to demand the use of group techniques. With respect to a group technique it has the advantage to examine each factor individually so as to avoid the problem of cancellation effects (two factors individually influential may belong to same group and have effects that partially cancel out).

## The elementary effects method

The method of the Elementary Effects (EE) is a simple but effective method to screen a few important input factors among the many that can be contained in a model.

The main idea underlying the method is due to Morris (Morris, 1991), who introduced the concept of elementary effect and proposed the construction of two sensitivity measures with the scope of determining which input factors may be considered to have effects which are (a) negligible, (b) linear and additive, or (c) non-linear or involved in interactions with other factors.

An elementary effect is defined as follows. Assume each model input $X_i$, $i$= 1, …, $k$, varies across $p$ selected levels in the space of the input factors. In other words the input space is discretized into a $p$-level grid Ω. For a given value of $\mathbf{X}$, the elementary effect of the $i$th input factor is defined as:

$$d_i(\mathbf{X}) = \frac{[y(X_1,..,X_{i-1},X_i+\Delta,X_{i+1},..,X_k) - y(\mathbf{X})]}{\Delta}$$

where Δ is a value in { 1/($p$-1), …. , 1-1/($p$-1)}, $p$ is the number of levels, $\mathbf{X} = (X_1, X_2,..., X_k)$ is any selected value in Ω such that the transformed point ($\mathbf{X}$ +

$\mathbf{e}_i \Delta$) is still in Ω for each index $i$=1,…, $k$, and $\mathbf{e}_i$ is a vector of zeros but with a unit as its $i$th component.

The finite distribution of elementary effects associated with the $i$th input factor, is obtained by randomly sampling different $\mathbf{X}$ from Ω, and is denoted by $F_i$, i.e. $d_i(\mathbf{X}) \sim F_i$. The number of elements of each $F_i$ is $p^{k-1}[p - \Delta(p-1)]$.

The sensitivity measures, μ and σ, proposed by Morris are respectively the mean and the standard deviation of the distribution $F_i$. The mean μ assesses the overall influence of the factor on the output. The standard deviation σ estimates the ensemble of the factor's higher order effects, i.e. non linear and/or due to interactions with other factors.

An intuitive explanation of the meaning of σ is the following. Assume for factor $X_i$ we get a high value of σ. Then the elementary effects relative to this factor differ notably from one another, implying that the values of the elementary effects are strongly affected by the choice of the sample points at which they are computed, i.e. by the choice of the other factors' values. In contrast a low value of $\sigma$ indicates very similar values of the elementary effects, implying that the effect of $X_i$ is almost independent of the values taken by the other factors.

Campolongo et al. (2007) proposed to replace the use of the mean μ with μ*, which is defined as the estimate of the mean of the distribution of the absolute values of the elementary effects that we denote with $G_i$, e.g. $|d_i(\mathbf{X})| \sim G_i$.

The use of μ* is convenient as it solves the problem of the Type II error (fail the identification of a factor of considerable influence on the model) to which the original measure μ can be exposed. Type II errors might occur when the distribution $F_i$ contains both positive and negative elements, which is when the model is non-monotonic. In this case, in computing their mean, some effects may cancel each other out, thus producing a low mean value even for an important factor.

To avoid Type II errors, Morris (1991) recommended to consider at the same time the values of μ and σ, as a factor with elementary effects of different signs would have a low value of μ but a considerable value of σ, and proposed a graphical representation in the (μ,σ) plane, allowing to interpret results by taking into account simultaneously the two sensitivity measures.

This is in general a valuable approach, but it may become problematic in the case of large models with multiple outputs. Moreover, in contrast to μ, μ* has the advantage that can be adjusted to work with a group of factors, i.e. to produce an overall sensitivity measure relative to a group (see below).

Campolongo et al. (2007) also showed that μ* is a good proxy of the total sensitivity index ST (Homma and Saltelli 1996, Saltelli et al. 2000), a variance based measure which quantifies the main effect plus all the interaction effects of a given factor. The total index is the sensitivity measure to use when the goal is that of identifying non-influential factors in a model (rather then prioritising the most influential ones). μ* is an effective substitute for the total index when the computational cost of ST is unaffordable.

In general, as the estimate of μ comes at no extra computational cost (the same amount of model executions is required), it is convenient to compute all the three statistics μ, σ, and μ*, so as to extract the maximum amount of sensitivity information. For instance, μ provides information on the signs of the effects that the factor has on the output. If μ is high, it implies not only that the factor has a large effect on the output but also that the sign of this effect is always the same. If, in contrast, μ is low while μ* is high, it means that the factor examined has effects of different signs depending on the point of the space at which the effect is computed. The sampling strategy to estimate the three statistics is described in detail in the next section.

**The sampling strategy and its optimization**

In order to estimate the sensitivity measures (i.e. the statistics of the $F_i$ and $G_i$ distributions), the design focuses on the problem of sampling a number $r$ of elementary effects from each $F_i$. As the computation of each elementary effect requires two sample points, the simplest design would require $2r$ sample points for each input, for a total of $2rk$ model executions, $k$ being the number of input factors. Morris (1991) suggested a more efficient design that builds $r$ trajectories of $(k+1)$ points in the input space, each providing $k$ elementary effects, one per input factor, for a total of $r(k+1)$ sample points.

The idea to generate each trajectory is as it follows. A "base" value $\mathbf{x}^*$ for the vector $\mathbf{X}$ is randomly selected in the $p$-level grid $\Omega$. $\mathbf{x}^*$ is not part of the trajectory but it is

used to generate all the trajectory points, which are obtained from **x*** by increasing one or more of its *k* components by Δ. The first trajectory point, $\mathbf{x}^{(1)}$, is obtained by increasing by Δ one or more components of **x***, in such a way that $\mathbf{x}^{(1)}$ is still in Ω. The second trajectory point, $\mathbf{x}^{(2)}$, is generated from **x*** with the requirement that it differs from $\mathbf{x}^{(1)}$ in its *i*th component that has been either increased or decreased by Δ. In formula $x_i^{(2)} = x_i^{(1)} + \Delta$ or $x_i^{(2)} = x_i^{(1)} - \Delta$. The index *i* is randomly selected in the set $\{1,2, ..., k\}$. The third sampling point, $\mathbf{x}^{(3)}$, is generated from **x*** with the property that $\mathbf{x}^{(3)}$ differs from $\mathbf{x}^{(2)}$ for only one component *j*, for any $j \neq i$. It can be either $x_j^{(3)} = x_j^{(2)} + \Delta$ or $x_j^{(3)} = x_j^{(2)} - \Delta$. And so on until $\mathbf{x}^{(k+1)}$, which closes the trajectory. The design produces a trajectory of $(k+1)$ sampling points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, .., \mathbf{x}^{(k+1)}$ with the key properties that two consecutive points differ in only one component and that any component *i* of the "base vector" **x*** has been selected at least once to be increased by Δ. An example of trajectory for $k$=3 is illustrated in Figure 1.

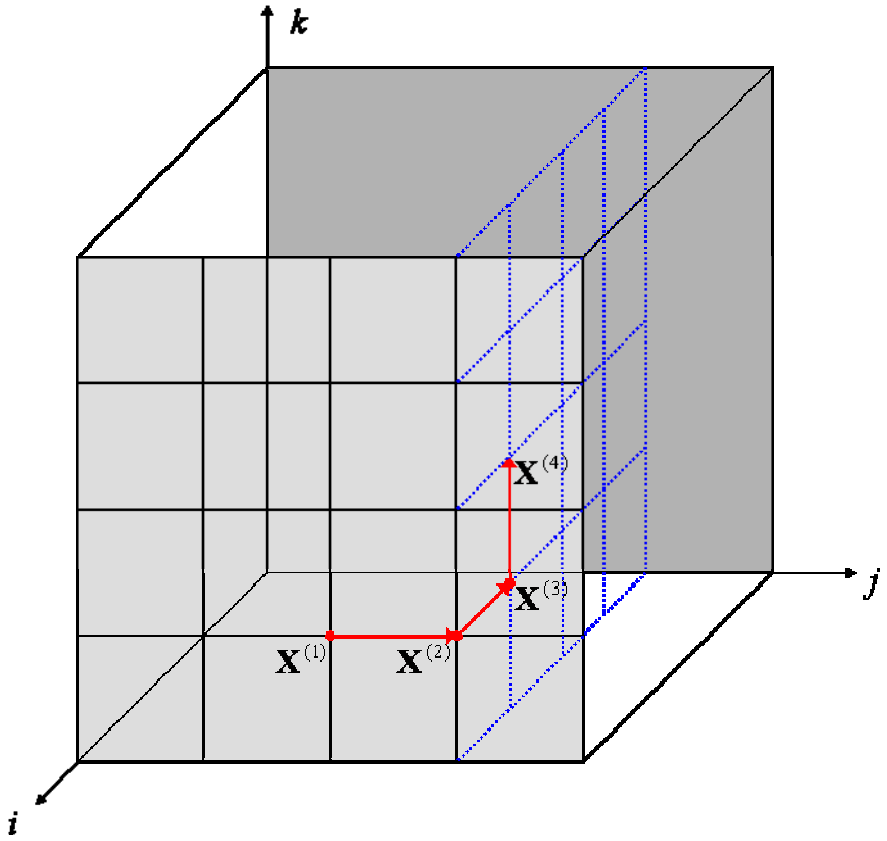


**Figure 1:** An example of trajectory in the input factor space when $k$=3

A technical scheme to generate trajectories with the required properties is the following. A trajectory can be seen in the form of a matrix, **B***, with dimension ($k$+1) x $k$, whose rows are the vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, .., \mathbf{x}^{(k+1)}$. To build **B***, the first step is the selection of a matrix **B**, whose dimensions are ($k$ +1) x $k$, with elements that are 0's and 1's and the key property that for every column index $j$, $j$=1, …, $k$, there are two rows of **B** that differ only in the $j$th entry. A convenient choice for **B** is a strictly lower triangular matrix of 1's.

The matrix **B'**, given by,

$$\mathbf{B}' = \mathbf{J}_{k+1,1}\ \mathbf{x}^* + \Delta\, \mathbf{B},$$

where

- $\mathbf{J}_{k+1,k}$ is a ($k$ +1) x $k$ matrix of 1's, and
- **x*** is a randomly chosen "base value" of **X**,

is a potential candidate to become the desired design matrix but it has the limit that the $k$ elementary effects it produces would not be randomly selected.

A randomized version of the sampling matrix is given by:

$$\mathbf{B}^* = (\mathbf{J}_{k+1,1}\ \mathbf{x}^* + (\Delta/2)\,[(2\,\mathbf{B} - \mathbf{J}_{k+1,k})\,\mathbf{D}^* + \mathbf{J}_{k+1,k}])\,\mathbf{P}^*,$$

where **D*** is a $k$-dimensional diagonal matrix in which each element is either +1 or -1 with equal probability, and **P*** is a $k$-by-$k$ random permutation matrix in which each column contains one element equal to 1, all others equal to 0, and no two columns have 1's in the same position. Read by rows, **P*** gives the order in which factors are moved; **D*** states whether the factors will increase or decrease their value along the trajectory. **B*** provides one elementary effect per input, which is randomly selected.

*Example*

Consider a model with two input factors taking values in the set {0,1/3,2/3,1}. In this case $k$=2, $p$=4, and Δ=2/3. Suppose the randomly generated **x***, **D*** and **P*** are

$$\mathbf{x}^* = (1/3,\ 1/3); \qquad \mathbf{D}^* = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}; \qquad \mathbf{P}^* = \mathbf{I}.$$

The matrix **B** is given by:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$ and for these values we get

$$(\Delta/2)[(2\mathbf{B} - \mathbf{J}_{k+1,k})\mathbf{D}^* + \mathbf{J}_{k+1,k}] = \begin{bmatrix} 0 & \Delta \\ \Delta & \Delta \\ \Delta & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2/3 \\ 2/3 & 2/3 \\ 2/3 & 0 \end{bmatrix}$$

and

$$\mathbf{B}^* = \begin{bmatrix} 1/3 & 1 \\ 1 & 1 \\ 1 & 1/3 \end{bmatrix},$$

so that $\mathbf{x}^{(1)} = (1/3, 1)$; $\mathbf{x}^{(2)} = (1, 1)$; $\mathbf{x}^{(3)} = (1, 1/3)$.

Figure 2 shows the resulting trajectory in the input space

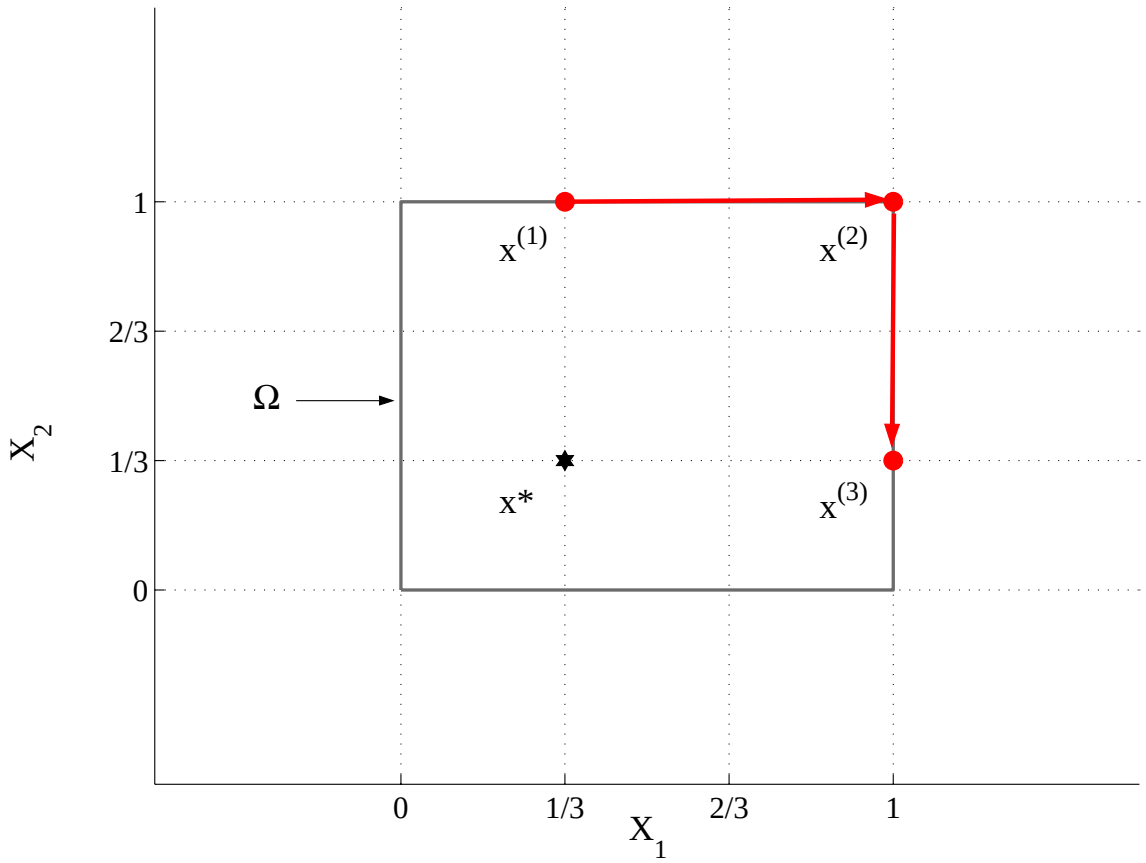


**Figure 2:** The trajectory obtained in the example

Campolongo et al., (2007) proposed an improvement of the sampling strategy just described that allows to achieving a better scanning of the input domain without increasing the number of model executions needed.

The idea is to select the *r* trajectories in such a way as to maximise their spread in the input space. The design starts by generating a high number of different Morris trajectories, *M*~500-1000, and then selects the subset of *r* (e.g. *r* = 10, 20) with the highest spread, where the concept of 'spread' is based on the following definition of 'distance.' The distance $d_{ml}$ between a couple of trajectories *m* and *l* is defined as:

$$d_{ml} = \begin{cases} \sum_{i=1}^{k+1}\sum_{j=1}^{k+1}\sqrt{\sum_{z=1}^{k}[X_i{}^m(z) - X_j{}^l(z)]^2} & \text{for } m \neq l \\ 0 & \text{otherwise} \end{cases}$$

where *k* is the number of input factors and $X_i^m(z)$ indicates the *z*-th coordinate of the *i*-th point of the *m*-th Morris trajectory. In other words $d_{ml}$ is the sum of the geometric distances between all the couples of points of the two fixed trajectories.

The best *r* trajectories out of *M* are selected by maximising the distance $d_{ml}$ among them. First we consider for each possible combination of *r* trajectories out of *M* the quantity $D^2$, which is the sum of the squared distances $d_{ml}$ between all possible pairs of trajectories belonging to the combination. For instance, if we select the combination 4, 6, 7, and 9 (i.e. *r*=4) out of the possible M={1,2,3,4,5,6,7,8,9,10}, we define $D_{4,6,7,9}$ as $D_{4,6,7,9} = \sqrt{d_{4,6}^2 + d_{4,7}^2 + d_{4,9}^2 + d_{6,7}^2 + d_{6,9}^2 + d_{7,9}^2}$. Then, we consider the combination with the highest value of *D*.

The revised sampling strategy optimises the scanning of the input space and is always to be preferred to the original one proposed by Morris (1991), as it offers a better scan of the input space without increasing the number of model evaluations required.

**The computation of the sensitivity measures**

The sampling strategy described above results in the construction of *r* trajectories in Ω. Each trajectory corresponds to (*k*+1) model executions and allows the computation of an elementary effect for each factor *i*, *i*=1,..,*k*.

If $\mathbf{x}^{(l)}$ and $\mathbf{x}^{(l+1)}$, with $l$ in the set $\{1,\ldots,k\}$, are two sampling points differing in their $i$th component, the elementary effect associated to the factor $i$ is either $d_i(\mathbf{x}^{(l)}) = [y(\mathbf{x}^{(l+1)}) - y(\mathbf{x}^{(l)})]/\Delta$, if the $i$th component of $\mathbf{x}^{(l)}$ has been increased by $\Delta$ or $d_i(\mathbf{x}^{(l)}) = [y(\mathbf{x}^{(l)}) - y(\mathbf{x}^{(l+1)})]/\Delta$, if the $i$th component of $\mathbf{x}^{(l)}$ has been decreased by $\Delta$.

Once $r$ elementary effects per input are available, the statistics $\mu$, $\mu^*$ and $\sigma$ relative to the distributions $F_i$ and $G_i$ can be computed by using the same estimators that would be used with independent random samples, as the $r$ elementary effects belong to different trajectories and are therefore independent.

A critical choice related to the implementation of the method is the choice of the parameters $p$ and $\Delta$. The choice of $p$ is strictly linked to the choice of $r$. If one considers a high value of $p$, thus producing a high number of possible levels to be explored, one is only in appearance augmenting the accuracy of the sampling. If this is not coupled with the choice of a high value of $r$, the effort will be wasted, since many possible levels will remain unexplored. In general, when the sampling size $r$ is small, it is likely that not all the possible factor levels are explored within the experiment. For instance, in the above example, if $r$ =1, the two factors never get the values 0 and 2/3. Previous experiments (Campolongo and Saltelli, 1997; Campolongo et al., 1999; Saltelli et al., 2000, p.367) demonstrated that the choice of $p$=4 and $r$=10 has produced valuable results.

If possible, it is convenient to chose for $p$ an even value and then set $\Delta$=p/[2(p-1)]. This choice has the advantage that the design sampling strategy guarantees equal-probability sampling from each $F_i$ (for details see Morris, 1991).

The top part of Figure 3 (grey arrows) shows that when $p$ = 4, the choice of $\Delta$ = p/[2(p-1)] = 2/3 guarantees that the 4 levels have equal probability to be selected. On the contrary, a choice of $\Delta$ = 1/3 would imply that the levels 1/3 and 2/3 are more often sampled. The two histograms below plot the empirical distributions obtained when generating $r$=20 trajectories for $p$ = 4 levels, in the case of $\Delta$ = 2/3 (left plot) and $\Delta$ = 1/3 (right plot).

The bottom part of the figure illustrates the case where an odd number of levels is considered ($p$ = 5). In this condition, no matter what value of $\Delta$ is chosen, it is

impossible to achieve equal probability for the elementary effects. In some cases, e.g. for $p = 5$ and $\Delta = ¾$, there are elementary effects which can be never sampled.

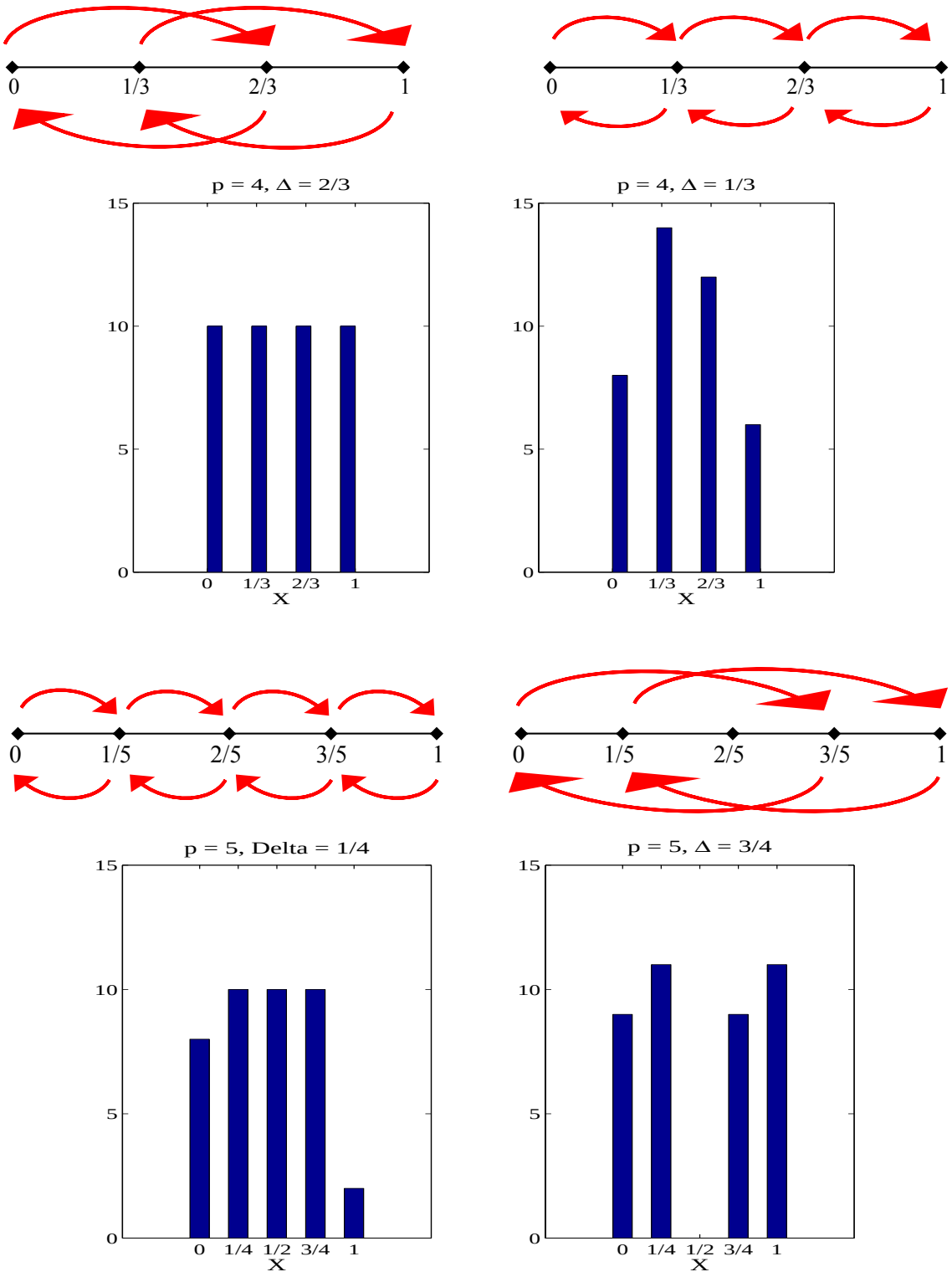


**Figure 3:** Empirical distribution obtained by sampling $r = 20$ trajectories and different values of $p$ and $\Delta$.

If a factor follows a uniform distribution, the levels are obtained by dividing in equal parts the interval in which each factor varies. If a factor follows other distributions than uniform, the sampling is carried out in the space of the quantiles of the distributions, which is a *k*-dimensional hyper-cube (each quantile varies in [0,1]). Then, the actual factor values are derived from its known statistical distribution (Campolongo et al., 1999).

Figure 4 shows the case where a factor X follows a normal standard distribution and the space of quantiles is investigated via six quantiles (q1, q2, …, q6). Using the inverse of the normal cumulative distribution function (black curve), the corresponding levels for X are obtained (L1, L2, …, L6).

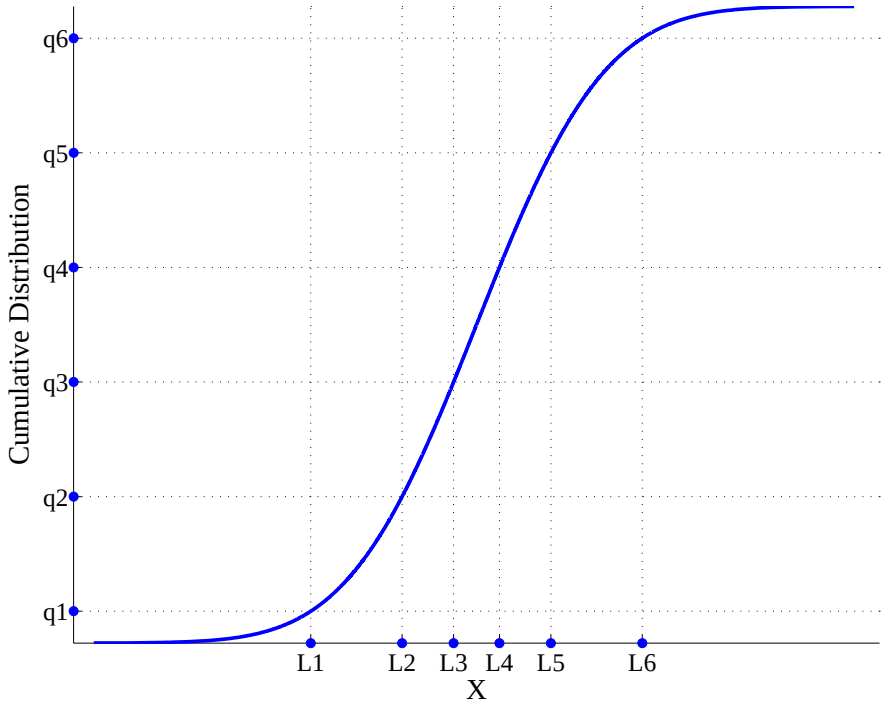


**Figure 4:** Sampling procedure for factors distributed as a standard normal.

At this point, it is worth spending a few words on the role of Δ. If Δ is chosen to be equal for all input factors, its role in the definition of the elementary effect for sensitivity purposes becomes irrelevant. $\Delta^{-1}$ is just a constant multiplying each elementary effect that does not affect the sensitivity analysis results. A question may arise when an input factor needs to be rescaled as its original distribution is not uniform between 0 and 1: *Should the Δ be rescaled so to represent the actual sampling step in the factor range of variation, or should it remain the same so to*

*represent the sampling step between [0,1] equal for each factor?* The following example answers this question. Assume the output y is the simple function $y = x_1 + x_2$, with $x_1$ and $x_2$ uniformly distributed respectively in [0,1] and [0,10]. In our sensitivity results we would like to obtain that input factor $x_2$ is much more important than input factor $x_1$, since a variation in its value affects y much more than a variation in $x_1$. Here we assume 4 levels, i.e. p=4, and r =1, i.e. we compute just one elementary effect for input and consider these as the final sensitivity measures. Assume that, following the standard Morris sampling strategy, we randomly generate the following 3 = r (k +1) points: (0,1/3); (0, 2/3); (1/3, 2/3). The reader experienced with the method can easily verify that this is a typical sample. The elementary effects relative to each input are then computed after rescaling input values for $x_2$ to be in the interval [0,10]:

$$d_1(\mathbf{X}) = \frac{y(1/3,20/3) - y(0,20/3)}{\Delta} = \frac{1/3}{\Delta},$$

$$d_2(\mathbf{X}) = \frac{y(0,20/3) - y(0,10/3)}{\Delta} = \frac{10/3}{\Delta}.$$

If in the computation of $d_2(\mathbf{X})$ the sampling step $\Delta$ is rescaled and set equal to 10/3, than both elementary effects result to be equal to 1, stating that both factors are to be considered as equally influent on the value of y. If instead $\Delta$ is left equal to 1/3 for both factors, independently of their actual ranges of variation, we would obtain a sensitivity measure for $x_2$ ten times higher than that of $x_1$. This second result is more sensible. In this way we are considering a sensitivity measure capable of taking into account not only the specifications of the model function, but also the effect of the statistical distributions assumed for the inputs.

In general, whatever type of distribution is considered, $\Delta$ should represent the sampling step between [0,1], i.e. in case of non-uniform distributions it should represent the variation in the quantiles of the factors.

**Working with groups**

The EE method presented above can also be extended to work with groups. When working with groups, the idea is to move all factors of the same group simultaneously. In the original definition given by Morris, the elementary effect is obtained by subtracting the function evaluated at **X** from that evaluated after incrementing one

factor. This definition can not be extended straightforwardly to cases in which more than one factor are moved at the same time, as two factors may have been changed in opposite directions, i.e. one increased and one decreased by Δ. By contrast, using μ* overcomes this problem, as the focus is not on the elementary effect itself but on its absolute value.

For a group $\mathbf{u} = (X_{i_1}, X_{i_2})$, the absolute elementary effect in point **X** is:

$$\left|d_{\mathbf{u}}(\mathbf{X})\right| = \frac{\left|y(\widetilde{\mathbf{X}}) - y(\mathbf{X})\right|}{\Delta}$$

where $\mathbf{X}$ is any selected value in Ω such that the transformed point $\widetilde{\mathbf{X}}$ is still in Ω, and each of the components $\widetilde{X}_{i_1}, \widetilde{X}_{i_2}$ have been either increased or decreased by Δ with respect to $X_{i_1}, X_{i_2}$. More details on the use of EE for groups can be found in Campolongo et al., (2007).

The sampling strategy described above needs to be slightly modified. It is first necessary to consider a matrix **G** describing how factors are apportioned into groups. This matrix is defined as it follows: the element G(*i, j*) of matrix **G** equals 1 if factor *i* belongs to group *j*; otherwise G(*i, j*)=0. Suppose *g* is the number of groups in the experiment. In this case the matrix of trajectories **B*** has dimensions (*g*+1) x *k*, since all the factors in a group move together. The matrix of trajectories **B*** can be built considering a lower triangular matrix **B** whose dimensions are (*k* +1) x *g* and setting:

$$\mathbf{B}^* = \mathbf{J}_{g+1,1}\ \mathbf{x}^* + (\Delta/2)\ [(2*\mathbf{B}\ (\mathbf{G} \times \mathbf{P}^*)^{T} - \mathbf{J}_{g+1,\,k})\ \mathbf{D}^* + \mathbf{J}_{g+1,\,k}],$$

where $\mathbf{J}_{i,j}$ is a matrix of ones with dimensions (*i* x *j*); **D*** is matrix (*k* x *k*) describing if the factors increase or decrease their values; and **P*** describes the order in which the groups move, and its dimensions are *g* x *g*. The following example illustrates how to handle groups of factors.

*Example*

Consider 3 factors X1, X2, and X3, uniformly distributed on [0, 1] and apportion them into a number of groups *g*= 2. The first group (G1) contains only factor X1; the second one (G2) includes the other two factors. The matrix **G** is defined as:

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Consider an experiment with p = 4 levels and to choose Δ=2/3 and suppose to get the following matrices for **x***, **D***, and **P***:

$$\mathbf{x}^* = [1/3,\ 1/3,\ 0];\ \mathbf{D}^* = \begin{bmatrix} 1 & & \\ & -1 & \\ & & 1 \end{bmatrix};\ \mathbf{P}^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The matrix $\Delta/2*[(2*\mathbf{B}\ (\mathbf{G}*\mathbf{P}*)^T - \mathbf{J}_{g+1,\,k})\ \mathbf{D}* + \mathbf{J}_{g+1,\,k}]$,equals in this case $\begin{bmatrix} 0 & \Delta & 0 \\ \Delta & \Delta & 0 \\ \Delta & 0 & \Delta \end{bmatrix}$,

which makes clear that $X_1$ (i.e. G1) moves first and increases its value, and then the factors in G2 change their values in opposite directions ($X_2$ decreases and $X_3$ increases). The final matrix is then:

$$\mathbf{B}^* = \begin{bmatrix} 1/3 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1/3 & 2/3 \end{bmatrix}.$$

Note that the two factors in the same group take also values into different levels.

**The EE method step by step**

In this section we show how to implement in practice the EE method illustrated above. The method is tested on the analytical *g*-function due to Sobol' (1993):

$$g = \prod_{i=1}^{k} g_i(X_i)$$

where $g_i(X_i) = \dfrac{|4X_i - 2| + a_i}{1 + a_i}$, and the $a_i$ are parameters, such that $a_i \geq 0$.

This function is widely used as a test function in sensitivity analysis because it is a very difficult one: it is strongly non-linear and non-monotonic, and all its interaction terms are non-zero by definition. The values of the $a_i$ determine the relative importance of the $X_i$, as they determine the range of variation of each $g_i(X_i)$:

$$1 - \frac{1}{1 + a_i} \leq g_i(X_i) \leq 1 + \frac{1}{1 + a_i}.$$

Thus, the higher the $a_i$ value, the lower the importance of the Xi variable. This is also shown by Figure 5, which illustrates the behaviour of $g_i(X_i)$ as a function of Xi for the values of $a_i$ =0.9, $a_i$ =9, and $a_i$ =99.

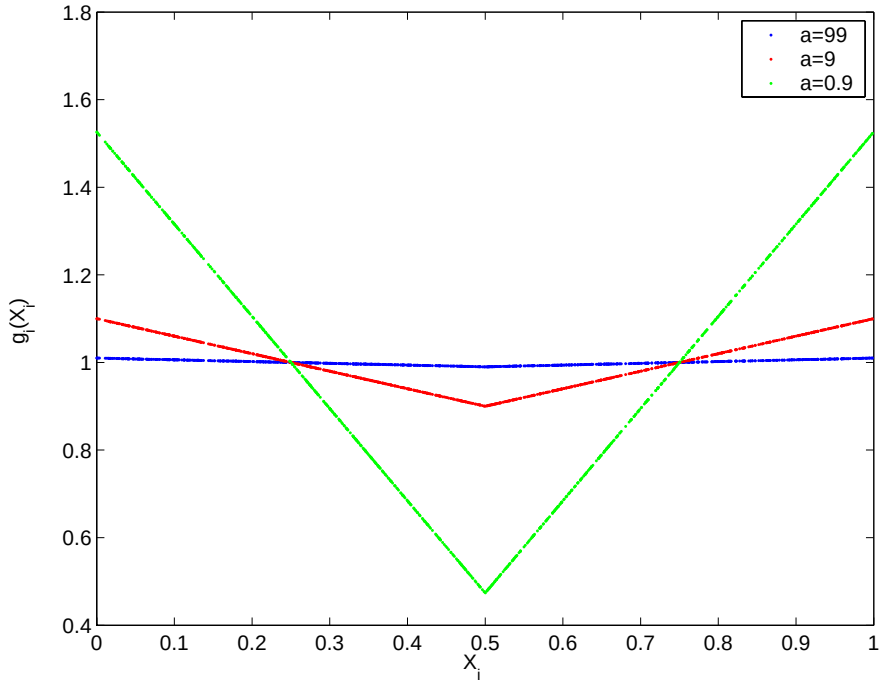


**Figure 5:** $g_i(X_i)$ as a function of Xi for $a_i$ =0.9, $a_i$ =9, and $a_i$ =99.

In our example we assume $k$=6 and we define the $a_i$ parameter values as:

| ***a*(1)** | ***a*(2)** | ***a*(3)** | ***a*(4)** | ***a*(5)** | ***a*(6)** |
|---|---|---|---|---|---|
| 78 | 12 | 0,5 | 2 | 97 | 33 |

**Table 1:** Values of the parameters of the g-function for each input.

We now perform the sensitivity analysis by making use of the sensitivity method described above. The uncertainty in the input variables is modeled by assuming that all the Xi are uniformly distributed in the 6-dimensional unit cube (i.e. $X_i \sim U[0,1]$, all $i$).

The optimized strategy described above is used to estimate the sensitivity measures μ, μ* and σ. $p$=4 levels, and Δ=2/3 are chosen; $r$=4 trajectories are employed, selected out of a bunch of 100. The sampled input matrix is presented in Table 2, (columns 2 to 7).

| | X1 | X2 | X3 | X4 | X5 | X6 | g |
|---|---|---|---|---|---|---|---|
| | 0 | 2/3 | 1 | 0 | 0 | 1/3 | **2,193** |
| | 0 | 2/3 | 1 | 0 | 0 | 1 | **2,280** |
| | 0 | 0 | 1 | 0 | 0 | 1 | **2,520** |
| **t1** | 2/3 | 0 | 1 | 0 | 0 | 1 | **2,478** |
| | 2/3 | 0 | 1 | 2/3 | 0 | 1 | **1,652** |
| | 2/3 | 0 | 1/3 | 2/3 | 0 | 1 | **0,771** |
| | 2/3 | 0 | 1/3 | 2/3 | 2/3 | 1 | **0,761** |
| | 0 | 1/3 | 1/3 | 1 | 1 | 2/3 | **1,024** |
| | 0 | 1 | 1/3 | 1 | 1 | 2/3 | **1,131** |
| | 0 | 1 | 1 | 1 | 1 | 2/3 | **2,424** |
| **t2** | 2/3 | 1 | 1 | 1 | 1 | 2/3 | **2,384** |
| | 2/3 | 1 | 1 | 1 | 1 | 0 | **2,478** |
| | 2/3 | 1 | 1 | 1 | 1/3 | 0 | **2,445** |
| | 2/3 | 1 | 1 | 1/3 | 1/3 | 0 | **1,630** |
| | 1 | 2/3 | 0 | 2/3 | 1 | 0 | **1,520** |
| | 1 | 2/3 | 0 | 0 | 1 | 0 | **2,280** |
| | 1/3 | 2/3 | 0 | 0 | 1 | 0 | **2,242** |
| **t3** | 1/3 | 2/3 | 0 | 0 | 1/3 | 0 | **2,212** |
| | 1/3 | 0 | 0 | 0 | 1/3 | 0 | **2,445** |
| | 1/3 | 0 | 2/3 | 0 | 1/3 | 0 | **1,141** |
| | 1/3 | 0 | 2/3 | 0 | 1/3 | 2/3 | **1,097** |
| | 1 | 1/3 | 2/3 | 1 | 0 | 1/3 | **1,024** |
| | 1 | 1/3 | 2/3 | 1 | 0 | 1 | **1,064** |
| | 1 | 1/3 | 0 | 1 | 0 | 1 | **2,280** |
| **t4** | 1 | 1/3 | 0 | 1/3 | 0 | 1 | **1,520** |
| | 1 | 1/3 | 0 | 1/3 | 2/3 | 1 | **1,500** |
| | 1 | 1 | 0 | 1/3 | 2/3 | 1 | **1,657** |
| | 1/3 | 1 | 0 | 1/3 | 2/3 | 1 | **1,630** |

**Table 2:** Sampled trajectories and corresponding *g*-function values.

As an example, Table 3 shows how to estimate the sensitivity measures for factor X4.

| | $g(X_1,..,X_4+\Delta,...,X_6)$ | $g(X_1,..,X_4+\Delta,...,X_6)$ | $d_4(\mathbf{X})$ | $\lvert d_4(\mathbf{X})\rvert$ |
|---|---|---|---|---|
| **t1** | 1,652 | 2,478 | -1,239 | 1,239 |
| **t2** | 2,445 | 1,630 | 1,222 | 1,222 |
| **t3** | 1,520 | 2,280 | -1,140 | 1,140 |
| **t4** | 2,280 | 1,520 | 1,140 | 1,140 |

**Table 3:** Estimation of the distribution of the elementary effects (and their absolute values for factor X4.

The values of μ, μ* and σ are reported in Table 4. For the estimation of the standard deviation of the elementary effects of the j-th factor the unbiased estimator is applied. Results indicate that X3 and X4 are important factors, while factors X1, X5, and X6 can be regarded as non influential (see μ* values in the first column). The high values of σ for some factors also demonstrate that interactions are playing an important role in the model. Moreover, the low values of μ associated with high values of μ* indicate that factors have effects of oscillating signs.

| | **μ*** | **μ** | **σ** |
|---|---|---|---|
| $X_1$ | 0,056 | -0,006 | 0,064 |
| $X_2$ | 0,277 | -0,078 | 0,321 |
| $X_3$ | 1,760 | -0,130 | 2,049 |
| $X_4$ | 1,185 | -0,004 | 1,370 |
| $X_5$ | 0,035 | 0,012 | 0,041 |
| $X_6$ | 0,099 | -0,004 | 0,122 |

**Table 4:** Estimated sensitivity measures.
The measures are estimated using $r = 4$ trajectories.

Another way of reading results is presented in Figures 6 and 7, where the values of μ and μ* are plotted as a function of σ. The plots confirm that factors X2, X3, and X4

are the most important factors, and also that they have effect of opposite signs on the output. These two scatter-plots stress the gain obtained introducing the measure μ*.

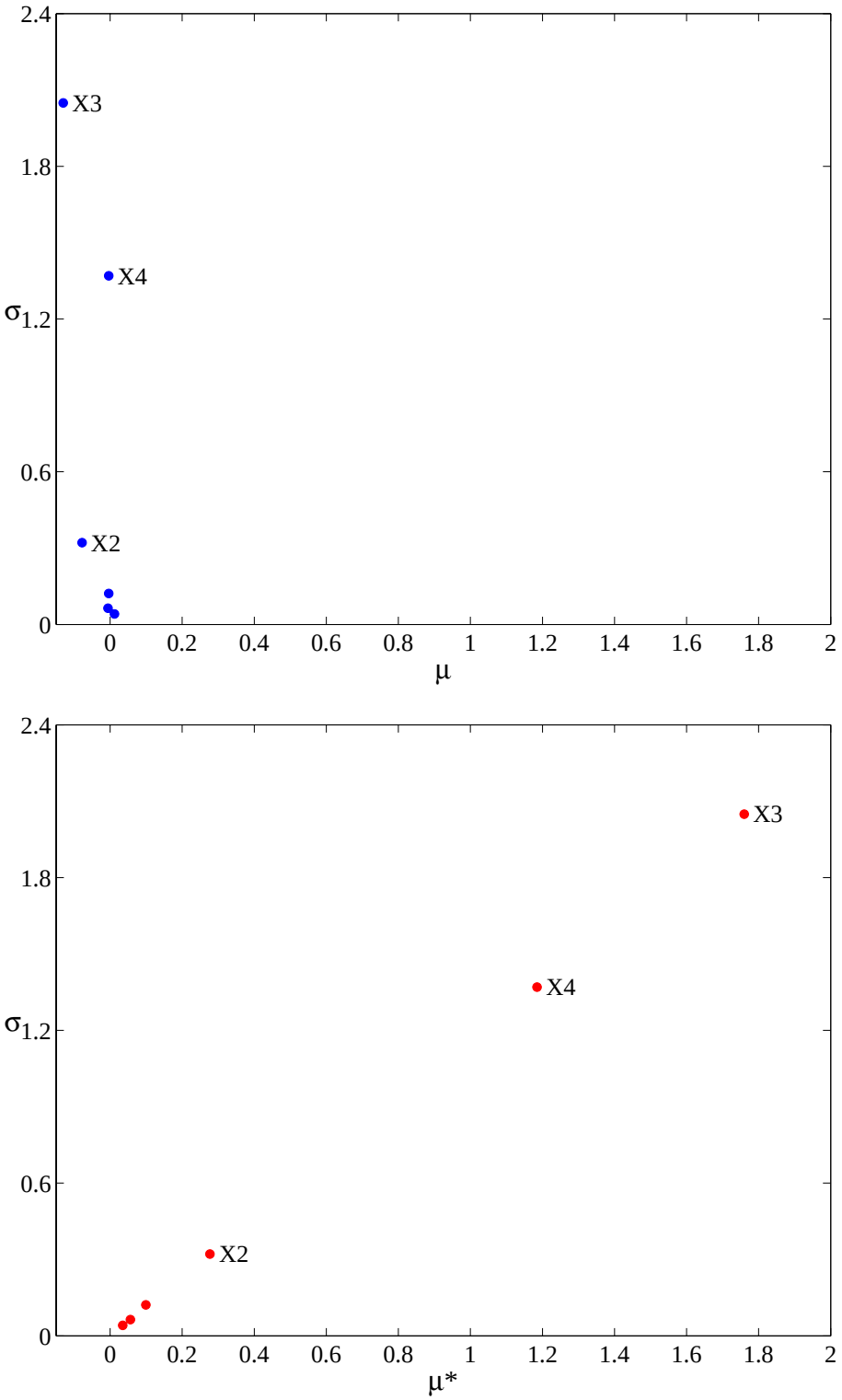


**Figure 6 and 7:** Scatterplots of {μ, σ} and {μ*, σ} for the *g*-function.

**Conclusions**

The elementary effects method illustrated in this work has proved to be effective in identifying the few important factors in a model that contains many factors, with a relatively small number of model evaluations.

The method is convenient when the number of factors is large, and/or the model execution time is such that the computational cost of more sophisticated techniques is excessive, but not so large to oblige the modeller to make use of grouping techniques.

The method is conceptually simple. It can be though of as an expansion of a derivative-based approach: when a single trajectory is considered, and the variations of input factors are small, it reduces to a derivative measure. Nevertheless, it overcomes the limitation of a local derivative-based approach as it attempts to explore the whole input space. This property is even reinforced when the improved sampling strategy due to Campolongo et al. (2007) is adopted.

The method is easy to be computed and has the good flexibility which is typical of the OAT approaches as it allows handling models that are unstable. When a model is unstable it risks to crash if executed on a set of inputs values rather different from the nominal values on which it was calibrated. As the OAT designs, which allows substituting a sample point where the model fails with a different one without changing the full design, the EE method, being based on trajectories independent from one another, allows substituting a trajectory on which the model execution fails with another one better handled by the model.

Last but not least, when needed, the method can be applied to groups, thus increasing the efficiency of the design.

## References


Campolongo, F., Cariboni, J., and Saltelli, A., (2007), "An effective screening design for sensitivity analysis of large models", *Environmental Modelling & Software,* in press.

Campolongo, F. and Saltelli, A., (1997), "Sensitivity Analysis of an environmental model: an application of different analysis methods", *Reliability Engineering and System Safety*, 57(1), 49-69.

Campolongo, F., Saltelli, A., Jensen, N. R.,Wilson, J. and Hjorth, J., (1999), "The role of multiphase chemistry in the oxidation of dimethylsulphide (DMS). A latitude dependent analysis", *Journal of Atmospheric Chemistry*, 32, 327-356.

Homma, T. and Saltelli, A., (1996), "Importance measures in global sensitivity analysis of model output", *Reliability Engineering and System t Safety*, 52, 1-17.

Morris, M. D., (1991), "Factorial Sampling Plans for Preliminary Computational Experiments", *Technometrics*, 33, 161-174.

Saltelli, A., Chan, K., and Scott, M., (eds.), (2000), *Handbook of sensitivity analysis*, John Wiley & Sons publishers, Probability and Statistics series.

Sobol' I.M., (1993), "Sensitivity Estimates for Nonlinear Mathematical Models", *Mathematical Modelling and Computational Experiments*, 1(4) 407-414.

European Commission

EUR22798 EN– DG Joint Research Centre, Institute for the Protection and Security of the Citizen
Sensitivity Analysis: How to Detect Important Factors in Large Models
Authors: CAMPOLONGO FRANCESCA, CARIBONI JESSICA

Luxembourg: Office for Official Publications of the European Communities
2007 – 21 pp. – 21 x 29,7 cm
EUR - Scientific and Technical Research series; 1018-5593

Abstract

In this work we propose a sensitivity analysis method which is effective in identifying the few important factors in a model that contains many factors, with a relatively small number of model evaluations.
The method is convenient when the number of factors is so large, and/or the model execution time is such, to make the computational cost of more sophisticated techniques excessive.
The method is conceptually simple and can be though as an expansion of a derivative-based approach, although it overcomes the limitation of a local derivative-based approach as it attempts to explore the whole input space.

**EUROPEAN COMMISSION**
DIRECTORATE-GENERAL
**Joint Research Centre**

The mission of the Joint Research Centre is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of European Union policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.